Digital Professionalism in Health and Care: Developing the Workforce, Building the Future
P. Scott et al. (Eds.)

doi:10.3233/SHTI220930

Preserving Privacy when Querying OMOP CDM Databases

Joao Rafael ALMEIDA[a,c,1], Joao Paulo BARRACA[b] and José Luís OLIVEIRA[a]
[a] *DETI/IEETA, University of Aveiro, Portugal*
[b] *IT, DETI, University of Aveiro, Portugal*
[c] *Department of Computation, University of A Coruña, Spain*

Abstract. Anonymisation is currently one of the biggest challenges when sharing sensitive personal information. Its importance depends largely on the application domain, but when dealing with health information, this becomes a more serious issue. A simpler approach to avoid inadequate disclosure is to ensure that all data that can be associated directly with an individual is removed from the original dataset. However, some studies have shown that simple anonymisation procedures can sometimes be reverted using specific patients' characteristics. In this work, we propose a secure architecture to share information from distributed databases without compromising the subjects' privacy. The anonymiser system was validated using the OMOP CDM data schema, which is widely adopted in observational research studies.

Keywords. Privacy preserving, Data anonymisation, k-Anonymity, l-Diversity, OMOP CDM, OHDSI

1. Introduction

Current approaches followed when sharing clinical data, simply try to avoid releasing sensitive information when publishing datasets. Their contents are often anonymized through processes that modify the original data, using data transformations that hide or remove subjects' identities, without degrading the data utility. However, anonymization simply based on the users' identity is limited, and there are still relevant challenges related to the privacy preservation of published data, namely on how to ensure the protection of data with smaller datasets, or on how to ensure resilience against future privacy threats [1].

The OMOP CDM (Observational Medical Outcomes Partnership Common Data Model) is the data schema used in the OHDSI community to store medical information in an interoperable format. This schema is person-centric and retains attributes that can be used to re-identify the subjects in the database. In this work, we propose the adoption of k-anonymity and l-diversity to increase data privacy when data owners decided to share a view of local databases.

[1] Corresponding Author: Joao Almeida, joao.rafael.almeida@ua.pt.

2. Methods

One of the most used techniques for anonymising data is *k*-anonymity. This privacy-preserving technique limits the information released, based on generalisation and suppression rules applied to the data concepts, as well as the number of repetitive elements for each equivalence class. An equivalence class is defined as a group of records that are indistinguishable from each other. A dataset to be compliant with *k*-anonymity needs to have an ambiguous map to at least *k* entities for each equivalence class[2]. Complementary to this technique, *l*-diversity aims to fill some gaps existent in the previous model, for instance when an equivalence class is unique or has a unique sensitive attribute. *L*-Diversity improves privacy by requiring that at least *l* "well-represented" values exist in the sensitive attributes [3]. In this work, we propose a system that implements these techniques to anonymise queries applied to OMOP CDM databases. The system was developed following a microkernel pattern, in which the plugins are defined as anonymisation algorithms, currently consisting in instantiations of *k*-anonymity and *l*-diversity. The classification of the data attributes was made using an ontology that also supports the generalisation and suppression rules required for these anonymisation techniques.

3. Preliminary results and next steps

The OMOP CDM databases are already classified as pseudo-anonymised databases. However, the procedure used is not robust and vulnerable to privacy attacks. In this work, we characterised the OMOP CDM fields depending on the information they correspond to. These attributes can be classified as sensitive, quasi-identifiers or key attributes. This work was validated using a synthetic patient population commonly used in the OHDSI community to validate cohorts and 5 real cohorts. The next steps for this work are the application of this tool using real data, aiming to enable data sharing instead of releasing statistical results about a medical study. This would increase the impact of medical studies if data owners adopt these techniques to protect their data.

Acknowledgment

This work has received support from the EU/EFPIA Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 806968. J.R.A. is funded by the FCT - Foundation for Science and Technology (national funds) under the grant SFRH/BD/147837/2019.

References

[1] Kaaniche N, Laurent M, Belguith S. Privacy enhancing technologies for solving the privacy personalization paradox: Taxonomy and survey. Journal of Network and Computer Applications. 2020;171:102807.

[2] Sweeney L. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 2002;10(05):557-70.

[3] Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD). 2007;1(1):3-es.